Identifying social norms using coordination games: Why does dictator game sharing vary?

Krupka, Erin L ; Weber, Roberto A

Abstract: We introduce an incentivized elicitation method for identifying social norms that uses simple coordination games. We demonstrate that concern for the norms we elicit and for money predict changes in behavior across several variants of the dictator game, including data from a novel experiment and from prior published laboratory studies, that are unaccounted for by most current theories of social preferences. Moreover, we find that the importance of social norm compliance and of monetary considerations is fairly constant across different experiments. This consistency allows prediction of treatment effects across experiments, and implies that subjects have a generally stable willingness to sacrifice money to take behaviors that are socially appropriate.

DOI: https://doi.org/10.1111/jeea.12006

Identifying social norms using coordination games: Why does dictator-game sharing vary?

Erin L. Krupka
School of Information, University of Michigan
and
IZA - Institute for the Study of Labor

Roberto A. Weber
Department of Economics
University of Zurich

February 20, 2012[*]

We introduce an incentivized elicitation method for identifying social norms that uses simple coordination games. We demonstrate that concern for the norms we elicit and for money predict changes in behavior across several variants of the dictator game, including data from a novel experiment and from prior published laboratory studies, that are unaccounted for by most current theories of social preferences. Moreover, we find that the importance of social norm compliance and of monetary considerations is fairly constant across different experiments. This consistency allows prediction of treatment effects across experiments, and implies that subjects have a generally stable willingness to sacrifice money to take behaviors that are socially appropriate.

JEL: C91, C72, D64
Keywords: social norms, experiment, dictator game

[*] The authors thank the Ford Foundation, IZA, and Carnegie Mellon University's Center for Behavioral Decision Research for financial support. We also gratefully acknowledge support from the research priority program at the University of Zurich "Foundations of Human Social Behavior." We thank Nicholas Bardsley, Colin Camerer, Cristina Bicchieri, Ted O'Donoghue, Tanga McDaniel, Stefano DellaVigna, Ulrike Malmendier, anonymous referees and participants at several workshops, seminars, and conferences for helpful comments and suggestions.

I. Introduction

Social norms have long been recognized as an important influence on behavior in social sciences such as social psychology (Sherif 1936; Cialdini et al. 1990) and sociology (Merton 1957; Coleman 1990). However, in economics social norms have received significant attention only relatively recently, mainly as a tool for explaining seemingly anomalous behavior such as involuntary unemployment (Akerlof 1980), conformity (Bernheim 1994), costly punishment (Fehr and Gächter 2000), tipping (Conlin et al. 2003), and macroeconomic phenomena, such as why consumption may track income even when wealth levels remain unaffected (Akerlof 2007).

One possible reason for the relative absence of social norms in economics is that they are difficult to measure or quantify, making it hard to predict the precise influence they will exert on behavior. As a result, social norms are usually incorporated into economic research as *post hoc* interpretations for behavior or outcomes that are otherwise difficult to explain (Fehr and Gächter 2000; Ostrom 2000), and they are identified primarily by measuring behaviors that are theoretically related to the norm (Fehr and Fischbacher 2004; Camerer and Fehr 2004). Because norms are usually studied indirectly in economics, they are rarely used to form precise predictions about behavior.[1]

In this paper, we aim to put the horse (norm) before the cart (behavior), by introducing a novel incentivized method for identifying social norms separately from behavior. We use this method to measure social norms in several economic choice contexts, and then use these elicited norms to predict behavior *a priori.* We do so in the context of other-regarding behavior in variants of the "dictator game," where recent laboratory experiments demonstrate that minor contextual features of a choice environment lead to substantially different choices and outcomes.[2] We show that such changes in behavior are entirely consistent with varying social norms and with a stable preference for complying with social norms.

---

[1] A few experimental studies manipulate the likely presence or strength of a social norm by varying features of a choice context (Krupka and Weber 2009; Andreoni and Bernheim 2009) and demonstrate resulting changes in behavior that are consistent with the influence of a norm or a preference for complying with a norm.

[2] This apparent "instability" in behavior has led some researchers to question the value of generalizing from such laboratory experiments to the field (Levitt and List 2007). Our work at least partially addresses this concern, by demonstrating how such behavior corresponds to varying and identifiable social norms. The sensitivity of behavior in the laboratory to the context of the experiment can be interpreted in a manner similar to how behavior in the field is sensitive to context and to varying social norms. For the laboratory data analyzed in this paper, we demonstrate that such sensitivity may be explained once varying social norms are identified. Therefore, the reason why someone might share in one dictator experiment but not in another very similar one might be the same as why one tips at a coffee shop but not at a fast-food restaurant, or in the U.S. but not in Europe.

Rather than attempting to develop a theory of norm compliance based on underlying preferences, as in prior research,[3] we start with the assumption that individuals care about behaving in a manner consistent with social norms. More precisely, we assume decision makers' utility is based on the money they obtain and on the degree to which their actions comply with social norms, in the form of taking actions generally viewed as socially appropriate and avoiding those viewed as socially inappropriate. We then show that these two considerations – combined with a novel, incentivized method for identifying social norms that uses coordination games – can explain significant behavioral variation in dictator games. Thus, our primary contribution is an empirical, rather than theoretical one. We present and demonstrate the usefulness of a novel method for norm elicitation by showing that the elicited social norms – when included as a component of utility in a conditional logit choice model – can account rather well for behavioral changes in experimental data. Moreover, we also show that the weights placed on money and norm compliance, in the estimated utility parameters in the conditional logit model, demonstrate a fairly constant willingness to trade off roughly $5 of wealth in order to take actions that are socially appropriate, rather than socially inappropriate. The stability of these preferences thus allows *a priori* predictions to be generated for new dictator-game contexts, once one obtains measures of the social appropriateness of different actions.

We begin by defining social norms and presenting a simple utility framework for understanding their potential influence on choice. We then demonstrate how one can use coordination games to identify the social norms that make up one source of utility. Using our utility framework, we show how the social norms we elicit from one set of individuals with the coordination games yield precise and testable predictions regarding the behavior of a new sample of participants, which we evaluate both with novel data from a new experiment and also using data from previously published experiments. We find that the observed sensitivity of behavior to

[3] For example, Andreoni and Bernheim (2009) explore situations in which one individual unilaterally shares wealth with another, and model a norm as the behavior that results from individuals caring about own wealth, intrinsically about fairness, and about how others perceive their concern for fairness. Their paper assumes an exogenously defined alternative (x^F), on which there is implicitly agreement that it is the "fair" action for the decision maker to take, and they follow prior research in assuming that the equal (50-50) division of wealth is a natural reference point. They then show why pooling may occur at this alternative (as well as at other alternatives, under changing conditions). Andreoni and Bernheim acknowledge that x^F may differ across contexts, and may thus account for varying behavior. This is similar to our main argument, that norms change across contexts, and can therefore account for changes in behavior, and suggests that our identification of what actions people agree upon as "appropriate" or "inappropriate" might provide an empirical basis for something like x^F in their model.

several surprising experimental treatments can be almost entirely explained by variations in social norms.

We choose to study behavior in dictator games primarily for two reasons. First, the simplicity and non-strategic nature of the dictator game make it easy to establish alternative environments in which we can hold constant important features of the choice faced by a decision maker (such as the set of possible payoffs and experimental subjects' understanding of how actions map into payoffs), while varying the context in a way that is likely to influence norms. The non-strategic nature of the dictator game also allows us to rule out the possibility that changes in behavior are due to changes in subjects' expectations about how opponents will behave, as would potentially be the case in a public goods, prisoner's dilemma, or trust games.

Second, a primary motivation in our research is to provide an interpretation for several recent experiments in economics (Dana et al. 2007; List 2007; Bardsley 2008; Lazear et al. 2012), all of which use variants of the dictator game. These experiments show that alternative treatments that make seemingly trivial or irrelevant changes to the choices available to a dictator nevertheless produce surprising changes in behavior. We provide at least one possible account for these changes. Indeed, part of the value in our approach is that it can provide an explanation for why behavior changes across dictator-game variants in a manner not adequately accounted for by most current theories of social preferences.[4] Thus, rather than adding to the literature on social preferences by conducting a "horse race" between different leading models, we purposely study simple decision contexts between which these models often fail to discriminate, and show that empirically-measured differences in social norms across these contexts correspond to observed differences in behavior.[5]

By applying our norm elicitation method to our own and others' dictator game data, we offer a unified interpretation of the behavioral changes observed across several experiments while simultaneously demonstrating the usefulness of the elicitation technique that we introduce. We also highlight the benefits of re-examining data from prior experiments to test novel theories

[4] In the Online Appendix, we consider several leading models of social preferences, and show that none of them can directly account for all the variation in behavior across the variants of the dictator game that we study here.

[5] Our approach might also serve as a valuable complement to existing theoretical approaches, providing a useful empirical input that can improve their ability to distinguish between environments that differ in social norms. For example, the norms that we empirically elicit could be incorporated into existing models of social norms, such as those by Lopez-Perez (2008) or Andreoni and Bernheim (2009) to provide a basis for why certain actions are considered "fair" or "compliant" with a norm, or to provide a basis for expectations in non-strategic settings (Battigali and Dufwenberg 2007).

and interpretations, as opposed to uniquely considering novel data generated solely for the purposes of a current test.[6]

We should also note that, while this paper is the first to introduce the coordination game-based elicitation method for identifying social norms, the approach has already been used in other papers. For example, Burks and Krupka (in press) apply the method in a real firm, to study how social norms regarding behavior toward clients differ between financial advisers and their supervisors, and show that mismatches in perceived social norms correlate with job dissatisfaction. In another example, Gächter, et al. (in press), use our method to elicit social norms in a setting where two workers provide costly effort to a firm, in a "gift-exchange" setting. They elicit norms to demonstrate that it is more socially appropriate to work hard when the other worker also exerts high effort, and compare the predictive ability of social norms to other social preferences.

The next section presents our method for identifying social norms. Sections III through V demonstrate the usefulness of this method for predicting and explaining behavior in non-strategic choice environments. Finally, in the conclusion, we discuss related work that validates the norm elicitation method.

II. Defining and Identifying Social Norms

Following Elster (1989), we note two important features of social norms. First, social norms generally prescribe or proscribe behaviors or actions, rather than outcomes. As Elster notes, "The simplest social norms are of the type: Do X, or: Don't do X." (p. 99). Allowing norms to govern actions, rather than outcomes, suggests that two actions that produce the same outcome, but differ in other respects, may be governed by different social norms. Second, the "social" element of norms requires that they be jointly recognized, or collectively perceived, by members of a population. These two features of social norms – that they apply to actions rather than outcomes and that they must be jointly recognized – are present in most researchers' definitions (Bettenhausen and Murnighan 1991; Fehr and Gächter 2000). For example, Ostrom

[6] While most experimental studies in economics (and psychology) limit their attention to a single dataset (usually generated by the authors), using additional, pre-existing data is valuable for the purposes of demonstrating robustness and generalizing findings beyond a particular experimental study. For other papers that employ this approach, see Camerer et al. (2004) and Hyndman et al. (in press).

(2000) defines social norms as "*shared understandings* about *actions* that are obligatory, permitted, or forbidden" (pp. 143-144, emphasis added).

Further, we distinguish norms regarding what one "ought" to do, or injunctive norms, from customs or actions that people regularly take, or descriptive norms. Both kinds of norms influence behavior (Cialdini et al. 1990; Krupka and Weber 2009). However, our focus here is on injunctive social norms, i.e., those described by Elster as prescribing what one "should do" or "should not do." As we will show, (injunctive) social norms concerning the appropriateness of behavior are sufficient for explaining a considerable amount of variation in other-regarding behavior.

Therefore, following the literature, we define (injunctive) social norms as collective perceptions, among members of a population, regarding the appropriateness of different behaviors. They are things that people in the population jointly recognize one should or should not do, and people who belong to the population expect others to be aware of and understand this agreement. The power of social norms comes both from the willingness of people within the population to punish (or reward) others' deviation from (or adherence to) them and from the experience of positive or negative emotions produced by one's own adherence or deviation from a social norm (Elster 1989; Fehr and Gächter 2000; Lopez-Perez 2008).

To formalize our definition, let $A = \{a_1, \ldots, a_K\}$ represent a set of K actions available to a decision maker. A social norm, $N(a_k) \in [-1, 1]$, is an empirically measurable collective judgment that assigns to each action a degree of appropriateness or inappropriateness. Therefore, we assume that if for an action, a_k, there is collective recognition that the action constitutes "appropriate," or socially prescribed, behavior, $N(a_k) > 0$, while if there is joint recognition that an action constitutes "inappropriate," or socially proscribed, behavior, $N(a_k) < 0$. Thus, consistent with the above definitions of social norms, $N(a_k)$ identifies the degree to which a specific action, a_k, is collectively perceived as one that should or should not be taken.

An important feature of the above definition, and where we depart from some prior work, is that a norm is not necessarily a binary classification, such that a particular action (the "norm", e.g., "tip 20%" or "the 50-50 split") should be taken, by assumption leaving all remaining actions as those (equally inappropriate) actions that should not be taken (Lopez-Perez 2008).[7] Instead,

[7] Such a definition is possible in our framework by, for example, assigning $N(a_k) = 1$ to only one action (the "norm") and letting all other actions have a constant value of $N(a_k) = 0$.

our definition of a social norm applies to the entire set of possible actions, and allows actions to vary in the degree to which they are perceived as appropriate. Thus, we can characterize a social norm by the *profile* of appropriateness ratings over all the actions available to a decision maker. For example, while there may be social agreement that it is always appropriate to arrive on time in many Western cultures, there may be some instances in which arriving 5 minutes late is less socially inappropriate (meeting friends at a bar) than others (arriving at a funeral). We will also see that representing social norms as a profile of appropriateness ratings over all actions is validated in our analysis of the experimental data. We show that even though the *most* socially appropriate behavior is the same across all the experiments we examine – share the experimental endowment equally (cf. Andreoni and Bernheim 2009) – we also find that *differences in the relative appropriateness* of the other actions exert an important influence on behavior.

To embed this definition of social norms in a simple utility framework – which will allow us to subsequently estimate the concern that individuals have for norm compliance, relative to money – we assume that a decision maker cares about both the monetary payoff produced by the selected action, $\pi(a_k)$, and the degree to which the action is collectively perceived as socially appropriate:

$$u(a_k) = V\big(\pi(a_k)\big) + \gamma N(a_k). \tag{1}$$

The function $V(\,)$ represents the value the individual places on the monetary payoff; we assume that this function is increasing in $\pi(a_k)$. The parameter $\gamma \geq 0$ represents the degree to which the individual cares about adhering to social norms. An individual entirely unconcerned with social norms ($\gamma = 0$) will always select the payoff-maximizing action. On the other hand, as γ increases, an individual will derive greater utility from selecting actions that are socially appropriate relative to the utility from those that are not.[8]

It follows directly from these preferences that behavior may change substantially across choice environments in which the sets of payoffs are identical, if the social norms change. Consider two choice environments, $A = \{a_1, a_2\}$ and $A' = \{a_1', a_2'\}$, such that $\pi(a_1) = \pi(a_1') > \pi(a_2) = \pi(a_2')$. Then, if there exist no social norms in either environment ($N(a_k) = N(a_k') = 0$,

[8] Other researchers have noted that individuals care heterogeneously about norm compliance (Ostrom 2000, Andreoni and Bernheim 2009). Such heterogeneity in pro-social concern is also common in most models of social preferences (Fehr and Schmidt 1999; Andreoni and Miller 2002). Cases in which $\gamma < 0$, which we do not explore here, might correspond to individuals who are anti-social, or derive utility from violating norms.

for $k = 1, 2$) the decision maker will choose a_1 in the first environment and a_1' in the second. However, if social norms differ between the two choice environments, the individual may select actions corresponding to different payoffs in the two environments. For example, if $N(a_1) = N(a_1') < N(a_2) < N(a_2')$, then for some values of γ a decision maker will select a_1 in the first environment and a_2' in the second environment. This is in spite of the fact that the most appropriate actions, a_2 and a_2', are analogous across the two contexts.

The above framework, while simple, presents a potentially useful approach for understanding how varying social norms might affect behavior even when choice environments are payoff-equivalent. It also provides a precise, and testable, relationship between the degree of social approval of actions ($N(a_k)$) and individuals' willingness to take those actions, provided one has a method for empirically measuring the "social appropriateness" of the different available actions.

In the rest of this paper, we predict and explain behavior using elicited measures of social appropriateness ($N(a_k)$). We first measure social appropriateness using a novel incentivized elicitation method. We elicit social norms over possible action choices across different contexts, from individuals who do not make choices in those contexts. We then observe how well the elicited social norms, when integrated into the above simple utility framework, explain the actual choices made by a separate group of individuals.

We measure the extent to which actions are socially appropriate or inappropriate by presenting respondents with a description of a choice environment, including all the possible available actions (i.e., $A = \{a_1, \ldots, a_K\}$). We ask respondents to judge the social appropriateness of *each* action – i.e., we elicit $N(a_k)$, for all a_k – on a four point scale that ranges over "very socially inappropriate", "somewhat socially inappropriate", "somewhat socially appropriate" to "very socially appropriate."[9] We provide respondents with incentives not to reveal their own personal preferences but instead to match the responses of others. Thus, respondents play a pure matching coordination game (Schelling 1960; Mehta et al. 1994) in which their goal is to

[9] The decision to have only four appropriateness categories was made after considering the tradeoff between having too few (in which case it would be harder to discriminate between degrees of appropriateness) and having too many (in which case it might be too difficult for subjects to match on the social norm, perhaps leading them to attempt to match using other focal principles). Further, we omitted the "neutral" category, as this would have been a focal point separate from the focal point stemming from the social norm.

anticipate the extent to which others will rate an action as socially appropriate or inappropriate, and to respond accordingly.

Because social norms reflect "collective perceptions," coordination games present a useful incentivized way to identify such socially held judgments. From a game-theoretic point of view, pure matching games such as the one we use in our experiment have many equilibria and nothing intrinsic to the payoffs of the game makes one equilibrium favored (or focal) over the others. Schelling (1960) theorized and Mehta et al. (1994) and Sugden (1995) demonstrated that prominence derived from common culture and shared experiences can create focal points. In our experiment, we allow collectively recognized social norms to create focal points in the matching game. Therefore, our elicitation method will yield a representation of a social norm if a) there is general social agreement that some actions are more or less socially appropriate, constituting the social norm, and if b) respondents attempting to tacitly match others' responses rely on such shared perceptions to help them do so.[10]

We begin by focusing on two payoff-identical variants of the dictator game. In Experiment 1 subjects see a description of *one* of these two choice environments, including all the possible choices available to the "dictator." From these subjects, who never actually play the dictator game described to them or see the other variant, we elicit social norms over actions in the described choice environment using the incentives we describe above. We then use the social norms elicited in Experiment 1 to predict behavioral changes across the two environments, and we test these predicted effects of social norms using data collected from a second, separate, group of subjects who make choices in one of the two environments (Experiment 2).

A second part of our analysis involves identifying social norms governing behavior in previously studied additional variants of the dictator game (Lazear et al. 2012; List 2007; Dana et al. 2006). Therefore, as part of Experiment 1, we also use our elicitation method to measure the degree of social appropriateness of different actions available in these particular experiments. We demonstrate that the identified social norms explain considerable variation across treatments in both our own experiment (Experiment 2) and across these previously studied experiments. We also use a conditional logic choice model to obtain estimates of the weights that individuals place

[10] Many previous researchers have noted the important relationship between social norms and equilibrium selection in games (Kandori 1991; Young 1998). Camerer and Fehr (2004) note that coordination games can be used with economic incentives to reveal shared understanding (see also Xiao and Houser 2011).

on complying with social norms (γ) and on monetary payoffs in several of these experiments, and show that a stable set of weights can explain a considerable amount of the variation in behavior across these experiments.

III. Identifying social norms in payoff-equivalent environments (Experiment 1)

Consider the following two choice environments. In a "standard" dictator game, a decision maker initially receives \$10 while another person receives \$0. The decision maker must then decide how much, between \$0 and \$10, in one-dollar increments, to *give* to the other person. In a "bully" variant of the game, the decision maker and other person both initially receive \$5 and the deciding individual can *give or take* any amount between \$0 and \$5, again in one-dollar increments, to or from the other person. Both choice environments offer the decision maker exactly the same 11 choices over final wealth allocations ranging from (\$10, \$0) to (\$0, \$10), but vary in the actions required to obtain those dollar allocations.[11]

While the two choice sets are identical in terms of final payoffs, they differ in contextual features of the actions required to achieve those payoffs. In the standard case, any outcome other than (\$10, \$0) involves "giving" money to the other person; in the bully variant all outcomes from (\$10, \$0) to (\$6, \$4) involve the decision maker "taking" from the other person. Therefore, it is possible that social norms governing the two sets of behaviors might differ, even though the resulting outcomes do not. In particular, we conjecture that social norms will differ over actions that involve "taking" vs. "giving," holding the resulting payoffs constant, in a manner that makes actions that involve "taking" less socially appropriate.

To identify social norms in the two choice environments, we applied our elicitation method to obtain ratings of the extent to which different actions in the two environments are collectively perceived as socially appropriate or inappropriate. Subjects providing the ratings saw only one of the two choice environments, and received incentives to match the modal response provided by others rating the same choice environment.

[11] Our experiment joins other research that examines the effect of varying initial endowment levels, such that dictators may "take" money from the recipient (Cox et al. 2007, Swope et al. 2008). Our "bully" variant differs from the dictator games with taking options studied by List (2007) and Bardsley (2008), which modify the standard dictator game by introducing *additional* taking options (see Section V).

A. Experimental Design for Experiment 1

We recruited 199 subjects from populations of experimental participants at Carnegie Mellon University, the University of Pittsburgh, and the University of Michigan.[12] Participants received $7 for showing up to the experiment and could earn additional money from a task in which they attempted to match others' appropriateness ratings. Subject payment in the matching task was not tied to the hypothetical dictator games about which they read.

The instructions (see Online Appendix) explained that subjects would read descriptions of different situations in which a person ("Individual A") faced a choice among several possible alternatives. For each situation, subjects were asked to rate the extent to which each alternative available to the person was "'socially appropriate' and 'consistent with moral or proper social behavior' or 'socially inappropriate' and 'inconsistent with moral or proper social behavior.'"

Participants then read, as an example, a hypothetical situation and were shown how they might indicate their ratings for each action in this situation.[13] After subjects were led through the example situation, but before they began to fill out the tables for the actual situations, they were told that one of the situations for which they were to provide appropriateness ratings would be selected at random at the end of the session, and that one of the possible action choices in this situation would also be randomly selected. If, for this action choice, the participant's appropriateness rating was the same as the modal response in the session, then that participant would receive an additional payment ($5 in Pittsburgh, $10 in Michigan) at the conclusion of the session. Thus, participants were incentivized to match the modal rating in their session, for each possible action.

Subjects then saw a description of *either* the standard or bully variant of the dictator game. Subjects in Experiment 1 never actually played this game, but only read about the situation and were asked to consider all of the actions that Individual A (the dictator) could take. In each session, only one of these two variants was used, meaning that no subject read descriptions of both the bully and standard choice contexts. The description of the situation

[12] Sessions conducted in Pittsburgh used 115 subjects and were conducted using pen and paper, while the sessions conducted at Michigan used 84 subjects and were conducted using the software z-tree (Fischbacher 2007). We find virtually no difference between the ratings obtained in Pittsburgh and in Michigan, despite there being differences in procedures (choices collected on paper in Pittsburgh vs. through a computer in Michigan; slightly different payoffs; sessions conducted years apart), indicating robustness of the social norms we elicit in these populations. We therefore pool across the two locations in the analyses.

[13] In the example, the decision maker found a wallet at a coffee shop and faced four alternatives: taking the wallet, asking others if the wallet belonged to them, leaving the wallet alone, or giving it to the store manager.

stated that the target individual (Individual A) was matched with another random and anonymous person (Individual B) and that both people would receive a "small participation fee" as well as any money produced by Individual A's choice.

The description then listed the eleven action choices available to Individual A. The labels associated with these action choices varied depending on which dictator game variant subjects were asked to consider (see Table 1). Subjects were also shown the monetary payments to each individual (A and B) produced by every listed action choice. For each possible action choice available to Individual A, a subject had to rate the choice as either "very socially inappropriate," "somewhat socially inappropriate," "somewhat socially appropriate," or "very socially appropriate," with the goal of matching this rating to the modal response in the session.

In the above manner, each subject provided social appropriateness ratings for all actions available in either the standard or bully variant of the dictator game. This yields our primary outcome measured in Experiment 1 – the "between-subjects" elicited ratings of social appropriateness, $N(a_k)$, for the bully and standard choice environments.

After rating all actions in either the bully or standard dictator variants, subjects then saw descriptions of either four (Pittsburgh) or five (Michigan) additional variants of the dictator game. Each situation corresponded to a variant of the dictator game used in previous experimental research (Lazear et al. 2012; List 2007; Dana et al. 2006). We discuss these variants in more detail in Section V.

After subjects indicated social appropriateness ratings in all choice scenarios, the experimenter randomly selected one scenario and one possible action choice in that scenario. The experimenter computed the modal response for that choice and privately informed subjects of whether or not their appropriateness rating matched the modal rating. Subjects were then paid privately in cash, receiving a $7 participation fee and an additional payment if they had selected the modal appropriateness rating for the selected scenario and action.

B. Results of Experiment 1

Recall that we conjectured that "taking" would generally be considered less socially appropriate than "giving," even when they produced identical outcomes. Therefore, we expected that, for those wealth allocations that left the dictator (Individual A) with more money than the

recipient, the corresponding actions would be generally considered less socially appropriate in the bully variant than in the standard variant of the dictator game.

We converted subjects' responses into numerical scores. A rating of "very socially inappropriate" received a score of -1, "somewhat socially inappropriate" a score of -1/3, "somewhat socially appropriate" a score of 1/3, and "very socially appropriate" a score of 1.[14] Table 1 presents subjects' social appropriateness ratings by condition. Each row corresponds to one possible action choice that Individual A could take and is also denoted by the final wealth distribution produced by that action choice, in the first column (payoff for A, payoff for B). For each of the two variants, the next several columns report first the action that the dictator had to take in order to obtain that wealth distribution, the mean of the social appropriateness ratings (ranging from complete agreement on "very socially inappropriate" (-1.0) to complete agreement on "very socially appropriate" (1.0)), and then the full distribution of responses. The final column reports the results of Wilcoxon rank-sum tests comparing the two distributions of responses.

Not surprisingly, the general pattern of social appropriateness ratings is the same across the two choice environments. There is substantial social agreement that the action that produces equal payoffs ($5, $5) is very socially appropriate in either environment. Further, maximizing A's own payoff and leaving the other person with nothing ($10, $0) is the most socially inappropriate action in either variant.[15]

However, as predicted, we also observe that actions involving "taking" are generally less appropriate than those involving "giving." For example, the action yielding payoffs ($10, $0) is less appropriate in the bully treatment than in the standard dictator game, and this difference is marginally statistically significant. Moreover, we observe even larger differences for outcomes in which the dictator obtains most, but not all, of the wealth. To obtain a payoff from $6 to $9 in the standard environment, the dictator must "give" to the other person while the bully environment requires the dictator to "take" from the other person to obtain the same payoffs. The

[14] We chose this particular scoring because it is intuitive (the least and most appropriate possible ratings receive scores of -1 and 1, respectively) and simple (possible ratings are evenly spaced over the -1 to 1 interval).
[15] Interestingly, actions that leave the recipient with more money (($4, $6) to ($0, $10)) produce less consensus. The modal and median responses lie between "very" and "somewhat" socially appropriate, but a significant proportion of respondents rate such behavior as socially inappropriate, and this proportion generally increases with other-regarding inequality. This might reflect the belief that it is socially inappropriate to be "too generous" – for example, when one gives a gift that is too expensive or when one attempts to tip a member of a profession that generally does not accept tips.

ratings confirm our expectation that "giving" is more socially appropriate than "taking." For every outcome from ($9, $1) to ($6, $4), the mean rating for the corresponding action is higher in the standard (giving) environment than in the bully (taking) environment, and these differences are all highly statistically significant.

Even in the cases where the ratings for the two environments diverge, subjects are still quite able to anticipate others' ratings – the modal response almost always receives over half of the responses. But what they agree upon often differs. For example, for the wealth allocation ($8, $2), the modal response in the standard environment for giving $2 to the other person is "somewhat inappropriate." But in the bully environment, where the same outcome involves taking $3 from the other person, there is social agreement that the action is "very inappropriate." Similarly, for the wealth allocation ($6, $4), the modal response in the standard environment is "somewhat appropriate," but in the bully environment it is "somewhat inappropriate."

C. Behavioral Predictions

The utility function in Equation 1 and the elicited norm ratings, $N(a_k)$, in Table 1 lead directly to two main predictions regarding how behavior will differ between the two environments.

Prediction 1: *More agents will select the action producing the equal-split ($5, $5) allocation in the bully environment than in the standard environment.*

Prediction 2: *Conditional on not selecting the action producing the equal-split ($5, $5) allocation, more agents will select the action producing the payoff-maximizing ($10, $0) allocation in the bully environment than in the standard environment.*

The two predictions are straightforward, and result from the fact that the greatest disparity in social appropriateness between the bully variant and the standard dictator games are for final allocations between ($6,$4) and ($9,$1), with such ratings always lower in the bully case.

The first prediction follows from the fact that the loss in social appropriateness from moving from the equal-split action to any action yielding a greater payoff for the decision maker is relatively greater in the bully treatment than in the standard environment. In the bully variant, every allocation that produces more wealth than ($5, $5) for the decision maker yields a lower

social appropriateness rating than in the standard dictator game.[16] Therefore, selecting the action corresponding to the equal split will be more attractive relative to every other feasible choice in the bully variant than in the standard game. More precisely, an individual's willingness to select the equal split action, $a_{(\$5,\$5)}$, over any other action that gives her more money, $a_k \in \{a_{(\$6,\$4)}, \ldots, a_{(\$10,\$0)}\}$, will depend on *(i)* her utility loss from foregoing the higher monetary payoff, $V(\$5) - V\big(\pi(a_k)\big) < 0$, on *(ii)* the degree to which the two actions differ in social appropriateness, $N\big(a_{(\$5,\$5)}\big) - N(a_k) > 0$, and on *(iii)* her concern for norm compliance, γ. Since we assume *(i)* and *(iii)* are invariant for the individual and across the choice contexts, the individual's willingness to choose the equal split action across the two contexts will depend on how $N\big(a_{(\$5,\$5)}\big) - N(a_k)$ differs between the bully and standard contexts. As Table 1 reveals, this difference in norm ratings is *always* larger in the bully context than in the standard dictator game. Thus, the equal-split allocation is relatively more attractive in the bully variant than in the standard game, relative to all other feasible choices.

The intuition behind the second prediction is similarly straightforward. Conditional on not selecting the equal-split action, implementing the payoff-maximizing action, $a_{(\$10,\$0)}$, is relatively more socially appropriate – compared to actions that produce payoffs between (\$9,\$1) and (\$6,\$4) – in the bully environment than in the standard one. That is, the difference between $N\big(a_{(\$10,\$0)}\big)$ and $N(a_k)$ is smaller in the bully context than in the standard dictator game, for all $a_k \in \big\{a_{(\$6,\$4)}, \ldots, a_{(\$9,\$1)}\big\}$.

Thus, in the bully environment, individuals are less likely to select something other than the equal split (Prediction 1), but if they do then they are more likely to take all the wealth (Prediction 2). Note that we do not generate a straightforward directional prediction regarding whether more or less will be shared in the bully and standard treatments. Instead, we have a slightly more complex set of behavioral predictions regarding how the distribution of action choices will vary with the choice context. A more concise way of summarizing our predictions is that more bimodal behavior will obtain in the bully treatment than in the standard treatment.

[16] Note that dictators should never allocate themselves less than half of the wealth because it produces both a lower monetary payoff and lower social appropriateness than choosing the action that produces (\$5, \$5).

IV. Evaluating predictions with behavioral choice data (Experiment 2)

To evaluate the accuracy of the above behavioral predictions, we conducted an experiment that placed a different set of subjects in one of the two choice environments for which we collected social appropriateness ratings in Experiment 1. These subjects, who had no knowledge of the coordination games used to elicit social norms in Experiment 1, played either the standard or bully version of the dictator game for monetary payoffs. The difference between the two environments was whether one subject in a pair received $10 and chose how much to give to the other subject (standard) or whether both subjects received $5 and one subject chose how much to give to or take from the other (bully). The possible set of final payoff allocations was identical in both environments.

A. Experimental Design for Experiment 2

Our experiment took place at the end of several large lecture classes at Carnegie Mellon University. We recruited participants by asking for up to 30 volunteers to remain after class for a 5-minute decision making experiment. Sessions consisted of between 16 and 30 participants. Participants received $2 for participating, in addition to any money from the allocation choices made in the dictator game variants described below.[17]

Once all non-participants left the classroom we randomly divided participants into two groups, seated in different areas of the room. One group (dictators) received instructions, which were also read aloud so that the other group (recipients) could hear the instructions.

In the *standard* dictator game, each dictator received a yellow envelope labeled "money for you" that contained ten $1 bills. The other group (recipients) received empty white envelopes labeled "money for other person." Instructions were read aloud describing the choice, in which dictators would make a (double-blind) anonymous decision of how much of the $10 in their envelope to share with the paired recipient. Dictators made allocation decisions by distributing money between the two envelopes, similarly to Hoffman, et al., (1994).

After instructions were read aloud, one experimenter collected the empty white envelopes from recipients and waited by the door to the hallway outside the room. Dictators exited the room one at a time, and each dictator received, from the experimenter, one recipients' empty

[17] We recruited from science and math classes, which tend to be the largest classes outside of economics and psychology at Carnegie Mellon. While we did not collect gender information, our sample was predominantly male.

white envelope prior to exiting (the dictator already had his or her yellow envelope in hand). Outside the room, in the hallway, dictators found a large sealed box with an open slit at the top. As described in the instructions, the dictator privately allocated money between the two envelopes, placed the white envelope labeled "money for other person" inside the box, and left with whatever remained in the yellow envelope. This procedure allowed individual decisions to be anonymous.[18] This concluded the experiment for the dictator.

Once all dictators had left, one of the experimenters brought the box back into the classroom, where the recipients had been instructed to form a line. As each recipient stepped up to a table, one at a time, an experimenter opened a white envelope, counted the number of $1 bills aloud, and handed the bills to the recipient. The other experimenter recorded the amount received by the recipient. This concluded the experiment.[19]

In the *bully* variant of the dictator game, procedures were identical except that the two envelopes handed out at the beginning of the experiment each contained five $1 bills. The instructions informed dictators that they would be able to give up to $5 to or take up to $5 from the other person.

B. Results of Experiment 2

Figure 1A presents the results. There were 52 dictators (104 subjects) in the standard treatment and 54 dictators (108 subjects) in the bully treatment. The mean amount allocated to the recipient was $2.46 in the standard game and $3.11 in the bully treatment. The results in the standard treatment are similar to those in other dictator games: subjects share about 20-25% of the endowment and most dictators share some money (see Camerer 2003, Engel 2010).[20]

Table 2 presents statistical tests of the changes in behavior across the two treatments. We include a control variable for the size of the class from which students were recruited – class size

[18] The box was placed in such a way that the experimenter standing at the door could see part of the back of the person standing at the box (but not enough to be able to determine whether the subject was reallocating money between the envelopes or when the envelope was being placed in the box). The experimenter could observe the subject departing from the box area, which allowed the experimenter to know when to send the next dictator out of the classroom. This minimal observation sufficed to prevent subjects from being able to open the box undetected.
[19] For accounting purposes, and to maintain anonymity of actions, dictators signed a sheet stating that they received a $2 participation fee as well as $10 to allocate between themselves and another participant. Recipients conversely signed a sheet stating that they received $2 and may have also received some money from another participant.
[20] The amount shared is higher than in other experiments with this high level of anonymity (Hoffman et al. 1994). However, the social distance between dictators and recipients in our experiment is probably lower than in typical studies, as they are classmates, and social distance is negatively related to sharing (Bohnet and Frey 1999).

ranged from 87 to 184 – since this is potentially a measure of social distance (Bohnet and Frey 1999). As expected, class size is generally negatively related to amount shared. The first model demonstrates that more is shared with the recipient in the bully treatment, relative to the standard. The next two columns test the two behavioral predictions based on the norm elicitation results in Experiment 1.

The first prediction was that more participants would select the action corresponding to the equal-split allocation ($5, $5) in the bully treatment than in the standard treatment. Figure 1A reveals strong support for this prediction. If we exclude those subjects who shared more than $5,[21] then in the standard condition 8 of 48 participants (17 percent) gave $5 to the recipient. In the bully treatment, however, the proportion is much higher: of 49 participants, 18 (37 percent) neither took from nor gave money to the recipient. As model 2 in Table 3 reveals, this difference in behavior is statistically significant ($p < 0.001$), providing support for Prediction 1.

The second prediction deals with what subjects do if they do not select the equal-split action. We predicted that, conditional on allocating less than $5 to the recipient, more dictators would leave the recipient with $0 in the bully variant than in the standard game. In the standard game, 40 participants gave less than $5 to the recipient, and of these 16 (40 percent) gave $0. In the bully variant, 31 participants took money from the recipient, and of these 16 (52 percent) left the recipient with $0. As we predicted, the percentage is higher in the bully variant than in the standard, and model 3 in Table 3 reveals this difference to be statistically significant ($p = 0.03$).

The net result of the two predicted effects is that far fewer subjects leave recipients with amounts from $1 and $4 in the bully variant than in the standard dictator game. In the standard game, 24 of 52 subjects (46 percent) share an amount from $1 to $4. But in the bully variant, this proportion is much lower (15 of 54, or 28 percent). This difference is statistically significant in a non-parametric chi-square test ($\chi^2(1) = 3.85$, $p = 0.05$).

To further explore how well our elicited norms can account for the data, we estimate Equation 1 using the appropriateness ratings from Experiment 1 and the behavioral data from Experiment 2. We use a conditional (fixed-effects) logistic regression, in which the (binary) dependent variable is whether an action was selected and explanatory variables are characteristics of the possible action choices, in our case each action's social appropriateness and

[21] Such behavior is usually present, though rare, in most dictator experiments (Camerer 2003), and is inconsistent with most models of social preferences. Our results do not substantively change if we include those 9 participants in the analysis (4 in the standard game and 5 in the bully game).

monetary payoffs. For each alternative, we include the empirical mean social appropriateness rating ($N(a_k)$) elicited from Experiment 1 (see Table 1), which varies by treatment. The coefficient for appropriateness ratings provides an estimate of the weight on social appropriateness in Equation 1, or γ. To estimate the weight placed on monetary payoffs, we impose a linear restriction on $V(\,)$, such that, for any final payoff for the dictator, π, $V(\pi) = \beta\pi$. Thus, we estimate the weight, β, that individuals place on the money they receive from a particular choice. The resulting utility function is,

$$u(a_k) = \beta\pi(a_k) + \gamma N(a_k). \quad (2)$$

We use conditional logit to estimate the two weights, β and γ (McFadden, 1974). To account for the fact that our estimates of $N(a_k)$ are noisy, we bootstrap standard errors for the coefficients.[22]

We report the estimation results in model 1 of Table 3. The coefficient for the appropriateness rating is positive and statistically significant, signifying that the estimated appropriateness ratings have a positive relationship with behavior. The positive coefficient for β indicates that people care about their own monetary payoff. Moreover, the influence of social appropriateness on behavior is not just statistically significant, but also large in magnitude. The ratio, $2\gamma/\beta$, identifies how much money an individual is willing to sacrifice to take an action that is very socially appropriate ($N(a_k) = 1$) rather than one that is very socially inappropriate ($N(a_k) = -1$). This ratio indicates that subjects are willing to pay \$5.66 to comply with social norms. Model 2 interacts appropriateness ratings with the non-standard – e.g., bully – treatment. The resulting coefficient is statistically insignificant and its inclusion has little effect on the other coefficients. Thus, in making their choices, subjects in Experiment 2 appear to care about the social appropriateness elicited from subjects in Experiment 1, and they do so equally in the bully and standard treatments.

To get a sense of how well this simple utility framework qualitatively accounts for the data from Experiment 2, we calculated the predicted frequencies of choices in the two treatments, using the estimated parameters in model 1 of Table 3. These predicted choice frequencies are shown in Figure 1B. As the figure reveals, even though we impose a strong

[22] Specifically, for each model in Table 3 we obtain estimates of standard errors from 1000 replications with data randomly sampled (with replacement) from the original norm ratings for each scenario and from the choice data for each treatment, preserving the original sample sizes in each case. To explore robustness, we also show that the analysis in Table 3 is robust to using the median rating, rather than the mean (see Online Appendix).

linearity assumption on $V(\,)$, the estimated weights capture the general qualitative properties of the data in Figure 1A. For example, Figure 1B predicts, for the bully treatment relative to the standard game, a greater frequency of equal splits ($5), fewer choices that leave the recipient with between $1 and $4, and roughly equal proportions of choices in which the recipient receives nothing ($0), which are all consistent with the data.

To summarize, Experiment 2 demonstrates that behavior changes significantly across two payoff-equivalent choice environments, in a manner consistent with the *a priori* predictions derived from the elicited norm ratings from Experiment 1. We confirm that changes in behavior are accounted for by changes in the social appropriateness of seemingly identical – in terms of payoffs – actions. However, to further test our interpretation for varying behavior in dictator games, we now turn to testing the extent to which elicited social norms predict changes across a larger set of choice contexts, using data from dictator game variants studied in previous papers.

V. Re-analyzing previously collected dictator game data

In Experiment 1, after providing ratings of the social appropriateness of actions in either the standard or bully environment, all subjects also performed similar ratings for other variants of the dictator game, each corresponding to experimental treatments conducted in previous research. The task of providing ratings was presented in exactly the same format as for the standard and bully variants – subjects saw a list of the possible actions available to a hypothetical "Individual A" (the dictator) in that particular experimental treatment, and then attempted to match the ratings of social appropriateness for each possible action provided by other subjects in the session.[23] Participants' incentives were identical to those for the first variant (standard or bully) that they had encountered.

We now show that the elicited ratings are consistent with behavior in these experiments and that the surprising results produced by specific variants of the dictator game can be accounted for by changes in the social appropriateness of actions. We also explore the extent to which parameter estimates of β and γ, when combined with social appropriateness ratings for a new context, can predict surprising behavioral treatment effects, out of sample.

[23] The variants we studied differed in some cases between the sessions of Experiment 1 conducted in Pittsburgh and Michigan, as we make clear when this was the case. No feedback was provided in Experiment 1 until subjects had completed the entire experiment. That is, subjects first played a matching game for *either* the standard or bully variant of the dictator game, then completed matching games on other variants of the dictator game. After subjects completed the experiment, they received feedback about others' choices in one scenario.

A. Dictator game with a sorting option

Lazear et al. (2012) explored a variant of the dictator game in which subjects could opt to not play the game (by "passing"), in which case the dictator received $10 and the other participant received $0 without learning that a dictator game could have been played.[24] The introduction of this option, which replicates the payoffs produced from sharing nothing, has a strong effect on sharing, as described in Figure 2A, which pools data from Lazear et al.'s Experiments 1 and 2. Mean sharing decreases by about 50 percent when there is a costless sorting option, relative to the standard dictator game in which there is not one. This is largely the result of a majority of dictators in the sorting treatment selecting not to play the game.

To see why the sorting option is so frequently chosen (even among people who share positive amounts when no such option is available) and why behavior changes so significantly between the environments with and without sorting, we consider the ratings of social appropriateness given to actions in the two environments. Figure 3 presents the mean social appropriateness ratings for each action (represented in terms of the amount shared with the recipient, on the x-axis), both for the standard version of the dictator game and the variant with the additional ($10, $0) sorting option. The solid line presents the mean ratings (from Table 1) for the standard version of the dictator game; the dashed line presents the ratings from the sorting variant, conditional on the dictator choosing to play the game. The two lines are very close, indicating that social appropriateness ratings differ very little for actions in the dictator game based on whether subjects were required to play the game (standard) or had the option of not playing the game but then chose to play (Sorting (Play)).[25]

The square in Figure 3 corresponds to the mean rating given to the choice of taking the sorting option and *not* playing the game (Sorting (Opt Out)). This action implements a ($10, $0) payoff outcome, with the recipient remaining uninformed about the game. Thus, the resulting payoffs are identical to those from playing the game and keeping all the money. However, as the

[24] This kind of game was also studied by Dana et al. (2006), who use a ($9, $0) outside option, and by Broberg et al. (2007), who elicit prices to opt out of the game (see also DellaVigna et al. 2012). We focus on Lazear et al.'s experiment in which the payoffs from the outside option ($10, $0) are identical to payoffs attainable in the dictator game.

[25] In all comparisons between actions producing identical outcomes, conditional on the dictator playing the game, the differences in ratings are statistically insignificant. Recall that roughly half of the participants, when providing the (sorting, play) ratings, had previously rated the standard game while the other half had not (they had rated the bully variant). The ratings for these two groups do not differ statistically.

ratings reveal, opting out is considered far less socially inappropriate. The mean rating is -0.07 for Sorting (Opt Out) versus -0.80 for keeping all $10 in the standard game and -0.82 for choosing to play and keeping $10 in the sorting variant (Sorting(Play,$10)).[26]

To see how considerations of social appropriateness are likely to influence behavior between the two variants, consider the relatively high social appropriateness of Opt Out, relative to the other actions that yield a payoff of $10 for the dictator. Returning to Equation 1, this relatively high degree of social appropriateness makes this action desirable relative to other actions, as it produces the highest possible monetary payoff at a relatively low cost in terms of disutility from social inappropriateness. Thus, people who select to share positive amounts in the standard variant of the dictator game might prefer to opt out in the sorting variant, which provides a large monetary payoff for the dictator ($10) with less disutility from violating social norms than would obtaining the same payoff when playing the dictator game.

Using Lazear et al.'s data from Figure 2A, we can perform the same kind of conditional logit choice estimation that we did for Experiment 2, to obtain parameter estimates for γ and β.[27] Models 3 and 4 in Table 3 report the estimated parameters. As with the estimates for our Experiment 2, we again find that the coefficient for the social appropriateness ratings is positive and statistically significant. Also, model 4 reveals that interacting appropriateness rating with sorting treatment yields a statistically insignificant coefficient, indicating again that subjects appear to care about social appropriateness equally in the two treatments. Interestingly, the parameter estimates in model 3 are very similar to those in model 1. Indeed the estimate of the ratio, $2\gamma/\beta$, for model 3 is almost identical to the one in model 1 (5.68 vs. 5.66, respectively). Thus, our estimates indicate that subjects are willing to pay large and very similar amounts to comply with social norms in both our Experiment 2 and in the experiment by Lazear et al.

Figure 2B presents the predicted choice frequencies obtained from the coefficients estimated in model 3 of Table 3. As with Experiment 2, the predicted frequencies describe

[26] The differences in appropriateness ratings between Sorting(Opt Out) and keeping $10 when playing the game are highly statistically significant ($p < 0.001$) both for the standard game ($z = 9.12$) and for the sorting variant ($z = 11.47$). Interestingly, there is less agreement regarding the social appropriateness of "Opt Out" than there is for other choices in either variant. The modal rating is "somewhat inappropriate," but there are high frequencies of other responses. This suggests that social norms elicited using our method could also capture the extent to which agreement regarding appropriateness influences behavior, beyond just the average appropriateness we use here. Since we find that mean appropriateness ratings (and median ratings, as reported in the Online Appendix) do well in explaining behavior in dictator games, which is the focus of this paper, we do not pursue this issue further here.

[27] We pool choice data from Lazear et al.'s Experiments 1 and 2, which includes both the standard and sorting (with a $10 opt out payment) variants. This is also the data used in Figure 2A.

qualitative changes in the actual data rather well, and generally capture the behavioral effects of introducing the sorting treatment.

B. Dictator game with additional taking options

We also analyze two variants of the dictator game studied by List (2007). In a standard variant, dictators divided $5 between themselves and another participant, in $0.50 increments. Thus, other than the endowment and the range of possible allocations, this treatment corresponds to the standard $10 dictator game we studied earlier. In a "Take $1" variant, dictators could alternatively take $1 from the recipient, an option selected by many participants. The surprising result from List's experiment is that the introduction of the additional taking option causes a downward shift in the distribution of positive amounts shared, and dramatically decreases the frequency of people sharing half of the endowment. The data from List's experiment is displayed in Figure 4A. A similar result is observed by Bardsley (2008).

Figure 5 presents the mean ratings of social appropriateness for each action from the standard $5 dictator game studied by List and from the Take $1 variant.[28] On the middle and right sides of the graph, when the dictator leaves the recipient with $2.50 or more, the ratings are very similar. However, toward the left of the graph, the ratings differ substantially. For any amount shared with the recipient between $0 and $2, the action is more socially appropriate in the Take $1 variant, when the dictator could have taken money instead, than in the standard dictator game; this difference is significant for all amounts from $0 to $1.50 ($z > 2.45$, $p < 0.01$, in all four comparisons using rank-sum tests). Thus, giving small amounts to the recipient is more socially appropriate when one could have taken money instead. For example, while sharing nothing with the recipient ($10, $0) is rated as very socially inappropriate in the standard treatment (-0.75), the same action is rated much less harshly in the Take $1 treatment (-0.26).

The differences in Figure 5 can help explain why at least some individuals who share positive amounts in the standard dictator game may share less when the taking options are introduced. In particular, our appropriateness ratings identify a likely two-fold effect of

[28] The ratings for this experiment were collected only in sessions conducted in Michigan. In the earlier Pittsburgh sessions, we collected ratings on a stylized $10 version of the List experiment (for which choice data does not exist), and the analysis of this alternative data is reported in Krupka and Weber (2009). Additionally, the instructions in List's (2007) on-line appendix mistakenly included a take $0.50 option, which, in personal communication, he told us was not part of the experiment. Some of our ratings of social appropriateness included this additional alternative, which we omit from the analysis. Omitting sessions in which this additional option was included does not substantively change the results.

introducing the additional taking option. First, the presence of this option makes keeping all, or most, of the money less socially inappropriate, thus making these actions more attractive than in the standard dictator game. Second, the additional taking option may itself be attractive, as it gives dictators an opportunity to earn a higher payoff ($6) than in the standard dictator game.

Since we have choice data from List's experiment and corresponding social norm ratings from Experiment 1, we can again estimate the weights that subjects place on money and norm compliance using the conditional logit specification. We report the results in models 5 and 6 of Table 3. We find that subjects in List's experiment place more weight on money than did subjects in Experiment 2 and in the experiment by Lazear et al. However, the estimated weight on norm compliance is similar to those for the other experiments and is statistically significant. Moreover, as in earlier models, the interaction between social appropriateness and the Take $1 treatment is not statistically significant, indicating that norm compliance is fairly constant across the experimental conditions. In contrast with the two earlier experiments, where we estimated subjects to be willing to pay approximately $5.67 to take an action that is very socially inappropriate rather than one that is very socially appropriate, we estimated a smaller amount here, $2\gamma/\beta$ = $2.67.[29]

Figure 4B presents the predicted choice frequencies, estimated from model 5. The prediction captures several of the key aspects of the results in Figure 4A. In particular, while the modal behaviors in the standard game are sharing $0 and sharing half of the endowment ($2.50), the modal behaviors in the take $1 treatment are sharing $0 and taking money from the recipient, with the latter more frequent. Moreover, the choice whose frequency decreases the most when moving from the standard to the take $1 treatment is sharing money equally.

C. The stability of preferences across experiments

We have thus far shown that concern for compliance with the elicited social norms and for money can account for behavioral changes in three distinct experiments. An important additional question is how well a stable set of preferences – measured by constant values of β

[29] This highlights a potential scaling issue, since an obvious difference between List's experiment and the two other experiments is the size of the stakes in the dictator game ($5 vs. $10). Thus, the difference between "very socially appropriate" and "very socially inappropriate" might mean different things, in dollar terms, in games with different stake sizes. We discuss this issue further in concluding.

and γ can explain behavior across these experiments, and possibly others. We explore this important question in three ways.

First, we can use the parameters estimated in model 1 of Table 3, based only on behavior in our Experiment 2, to predict behavior in the Lazear et al. (2012), and List (2007) experiments. In the preceding sections, we used the actual behavior in each experiment to estimate a separate set of parameters for that experiment. However, a stronger test of the predictive power of elicited norms, when combined with the simple utility function in Equation 1, involves making predictions across experimental populations, i.e., using the parameters obtained from one experiment or set of experiments to predict behavior in another experiment.

We therefore took the elicited norm ratings for the Lazear, et al., and List experiments (from Figures 3 and 5, respectively), and used the estimated coefficients from model 1 in Table 3 ($\beta = 0.656, \gamma = 1.858$), which were obtained using only data from our Experiment 2, to generate predictions using the logistic choice model. We report the resulting predicted distributions in figures in the Online Appendix. In both cases, the predicted changes in behavior when comparing treatments are generally consistent with the observed patterns in the data. Thus, important treatment effects in both experiments are captured by our approach, even when we used parameter estimates obtained from another experiment.

Second, in models 7 and 8 in Table 3, we estimate the weights on monetary payoffs (β) and on compliance with social norms (γ) using the data pooled from all three experiments. Model 7 uses the pooled data to estimate these parameters under the assumption that they are equal across experiments. The results reveal roughly similar weights on the two considerations as in earlier models. Moreover, the ratio $2\gamma/\beta$ is equal to 4.95, implying that subjects are generally willing to pay around $5 to take actions that are socially appropriate instead of socially inappropriate. Model 8 introduces interaction terms for the two coefficients and the Lazear, et al., and List experiments. None of the four interaction coefficients is statistically significant, indicating that the same two parameters do a good job of explaining the pooled experimental data.

Finally, we can also study the extent to which our analysis allows us to predict behavior in an additional – slightly different – version of the dictator game, including another treatment that produces surprising results. In Experiment 1, we also asked subjects to provide ratings of the social appropriateness of different actions available to a dictator in a binary dictator game studied

by Dana, et al. (2006). This experiment included a "hidden information" treatment in which dictators were initially unaware of the payoff consequences of their actions for a recipient but could costlessly acquire this information, simply by clicking on a button. The surprising finding in this experiment was that many subjects opted not to acquire the information and, as a result, selfish behavior increased. In the Online Appendix, we show that the elicited norms, combined with Equation 1 and the coefficient estimates from Table 3 allow us to predict key properties of this treatment effect. In particular, our analysis predicts that modal behavior in the baseline (without hidden information) will be fair, that the proportion of such fair behavior declines substantially in the hidden information treatment, and that many subjects will choose to remain willfully ignorant. Thus, we provide yet another example of how the norms elicited for a particular context, when combined with the utility weights estimated from other contexts, can allow us to predict changes in behavior due to subtle treatment differences.

VII. Conclusion

Our work makes two important contributions to the study of social behavior in economics. First, we introduce a novel incentivized method for identifying social norms that uses coordination games. Second, we demonstrate that the elicited social appropriateness ratings, when combined with a simple utility framework in which decision makers care about norm compliance and about money, accurately predicts behavioral changes across several variants of the dictator game. We also find a relatively stable degree of concern for money and for social appropriateness.

Of course, it is important to be critical about the extent to which we actually measure social norms, and not something else. In coordination games, there are multiple equilibria and subjects could coordinate in ways that have nothing to do with norms. However, we observe subjects regularly agreeing in their ratings, in a way that varies both between different actions in each scenario and across scenarios, and also in a way that corresponds to our intuitions regarding changing social norms.[30] One might also argue that the responses correspond to what subjects

[30] Another way to test the extent to which our method identifies social norms is to see if it captures norms that can be externally validated. Another paper (Krupka et al. 2011), applies our elicitation instrument to measure norms of tipping and punctuality, how they vary by nationality, and how individuals recognize distinctions in norms across populations. We find support for the idea that our method measures social norms. For example, foreign-born students provide different appropriateness ratings when matching responses with US students than with people from

themselves would do if playing the game, or to what they think others will do. But, this interpretation suggests either that almost everyone believes that they would choose to split the wealth equally or that they believe others will do so, which is highly inconsistent with the actual data.[31] In order to predict behavior, we find it necessary to combine the elicited norm ratings with the self-interested motive in Equation 1, and find support for both motives in parameter estimation, meaning that the norm ratings alone do not simply track behavior independently.

It is also important to note some of the limitations in our results. For example, while our analysis shows that relatively stable weights on money and on social appropriateness can explain changes in behavior across experiments and treatments, there are some aspects of the data that these stable weights get wrong. For example, our analysis using weights derived form $10 dictator games to predict behavior in dictator games with $5 stakes generally predicts too much fair behavior (see Online Appendix). This is consistent with the intuition provided by an anonymous reviewer that, under significantly higher stakes, our predictions are very likely to require different sets of weights on money and social appropriateness. Therefore, an approach that more thoroughly identifies the relationship between social appropriateness and monetary considerations constitutes valuable future research.

Another limitation is that we study a relatively simple set of contexts – all variants of the dictator game. The relationship between social norms, as elicited with our method, and behavior may be more complex in other games, such as public goods games and trust games, where reciprocity and uncertainty may play a greater role. Nevertheless, at least for a first step, our admittedly simple approach has considerable value in explaining changing behavior across several experiments, and presents a useful starting point from which to build an improved understanding of social norms.

References:

Andreoni, J. and D. Bernheim. 2009. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica,* 77(5):1607-1636..

---

their home country, in a manner consistent with changing cross-national norms. Also, variation in ratings of social appropriateness across populations mirrors externally validated differences in social norms across populations.

[31] More directly, Krupka, Leider and Jiang (2011) use this elicitation technique and elicit social norms as well as beliefs about the distribution of actions actually taken in a double dictator game and in a Bertrand game. They find that subjects' appropriateness ratings over actions are not primarily driven by beliefs about others' likely actions.

Andreoni, J. and J. Miller. 2003. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70(2): 737-53.

Akerlof, G. 1980. "A Theory of Social Customs, of Which Unemployment May Be One Consequence." *Quarterly Journal of Economics,* 94: 149-75.

Akerlof, G. 2007. "The Missing Motivation in Macroeconomics." *The American Economic Review*, 97(1): 3-36.

Bardsley, N. 2008. "Dictator Game Giving: Altruism or Artefact." *Experimental Economics*, 11(2): 122-33.

Battigali, P. and M. Dufwenberg. 2007. "Guilt in Games." The American Economic Review, 97(2): 170-176.

Bernheim, D. 1994. "A Theory of Conformity." *Journal of Political Economy*, 102(5): 841-77.

Bettenhausen, K. and J. Murnighan. 1991. "The Development of an Intragroup Norm and the Effects of Interpersonal and Structural Challenges." *Administrative Science Quarterly,* 36: 20-35.

Bohnet, I. and B. Frey. 1999. "Social Distance and Other-Regarding Behavior in Dictator Games: Comment." *The American Economic Review*, 89(1): 335-39.

Broberg, T., T. Ellingsen and M. Johannesson. 2007. "Is Generosity Voluntary?" *Economics Letters*, 94(1): 32-37.

Burks and Krupka. in press. "A multi-method approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry." *Management Science*

Camerer, Colin F. 2003. *Behavioral Game Theory*, Princeton, New Jersey: Princeton University Press.

Camerer, C. and E. Fehr. 2004. "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists." Foundations of Human Sociality -- Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Ed. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis.

Camerer, C., T. Ho and J. Chong. 2004. "A Cognitive Hierarchy Model of Games." The Quarterly Journal of Economics, 119(3): 861-898.

Cialdini, R., R. Reno and C. Kallgren. 1990. "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places." *Journal of Personality and Social Psychology,* 58(6): 1015-26.

Coleman, J. 1990. Foundations of Social Theory. Belknap Press of Harvard University Press (Cambridge Mass.).

Conlin, M., M. Lynn and T. O'Donoghue. 2003. "The Norm of Restaurant Tipping." *Journal of Economic Behavior and Organization,* 52(3): 297-321.

Cox, J., D. Friedman and S. Gjerstad. 2007. "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior*, 59:17-45.

Dana, J., D. Cain and R. Dawes (2006). "What You Don't Know Won't Hurt Me: Costly (but quiet) Exit in Dictator Games." *Organizational Behavior and Human Decision Processes,* 100(2): 1993-201.

Dana, J., R. Weber and J. Kuang. 2007. "Exploiting Moral Wriggle Room: Behavior Inconsistent with a Preference for Fair Outcomes." *Economic Theory*, 33(1):67-80.

DellaVigna, S., J. List and U. Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics*, 127(1): 1-56.

Elster, J. 1989. "Social Norms and Economic Theory." *Journal of Economic Perspectives*, 3(4): 99-117.

Engel, C. 2010. "Dictator Games: A Meta Study." *Experimental Economics*, 14: 583-610.

Fehr, E. and U. Fischbacher. 2004. "Third-Party Punishment and Social Norms." *Evolution and Human Behavior,* 25(2): 63-88.

Fehr, E. and S. Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives,* 14:159-81.

Fehr, E. and K. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics*, 114(3): 817-68.

Fischbacher, U. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." Experimental Economics, 10(2): 171-178.

Gächter, S. D. Nosenzo, and M. Sefton. in press. "Peer effects in pro-social behavior: Social norms or social preferences." Journal of the European Economic Association.

Hoffman, E., K. McCabe, K. Shachat and V. Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior,*7: 346-80.

Hyndman, K. E. Özbay, A. Schotter and W. Ehrblatt. in press. "Convergence: An experimental study of teaching and learning in repeated games." *Journal of the European Economic Association.*

Kandori, M. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies*, 59: 62-80.

Krupka, E., S. Leider and M. Jiang. Unpublished Manuscript. "A Meeting of the Minds: Contracts and Social Norms".

Krupka, E., R.Weber and R. Croson. 2011. “When in Rome: Identifying Social Norms as a Group Phenomenon.” Unpublished manuscript.

Krupka, E. and R. A. Weber. 2008. "Identifying Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" IZA Discussion Paper 3860.

Krupka, E. and R.Weber. 2009. "The Focusing and Informational Effects of Norms on Pro-Social Behavior." *Journal of Economic Psychology,* Vol. 30: 307-20.

Lazear, E., U. Malmendier and R. Weber. 2012. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied*, 4(1): 136–63.

Levitt, S. and J. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Tell Us About the Real World?" *Journal of Economic Perspectives*, 21(2): 153-74.

List, J. 2007. “On the Interpretation of Giving in Dictator Games.” *Journal of Political Economy*, 115(3): 482-94.

López-Pérez, Raúl, 2008. "Aversion to norm-breaking: A model," Games and Economic Behavior, Elsevier, vol. 64(1): 237-267.

McFadden, Daniel. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, New York: Academic Press.

Mehta, J., C. Starmer and R. Sugden. 1994. "The Nature of Salience: An Experimental Investigation of Pure Coordination Games." *American Economic Review,* 84(3): 658-73.

Merton, R. 1957. *Social Theory and Social Structure*. Free Press (Glencoe, IL).

Ostrom, E. 2000. "Collective Action and the Evolution of Social Norms." *Journal of Economic Perspectives,* 14(3):137-58.

Schelling, T. 1960. The Strategy of Conflict. Cambridge, MA, Harvard University Press.

Sugden, R. 1995. "A Theory of Focal Points." *The Economic Journal*, 105(430):533-50.

Sherif, M. 1936. The Psychology of Social Norms. Harper and Row, New York.

Swope, K., J. Cadigan, P. Schmitt and R. Shupp. 2008. "Social Position and Distributive Justice." *Southern Economic Journal,* 74(3): 811-18.

Xiao, E. and D. Houser. 2011. "Classification of Natural Language Messages Using a Coordination Game." Experimental Economics, 14(1): 1-14.

Young, P. 1998. “Social Norms and Economic Welfare.” *European Economic Review*, 42: 821-30.

Table 1. Elicited norms ($N(a_k)$) for bully vs. standard dictator environments (data from Experiment 1)

Action (final wealth)	Standard (n = 107) (Initial wealth: $10, $0)						Bully (n = 92) (Initial wealth: $5, $5)						*rank-sum test (z)*
	Action	**Mean**	- -	-	+	+ +	**Action**	**Mean**	- -	-	+	+ +	
$10, $0	"Give $0"	**-0.80**	82%	10%	3%	5%	"Take $5"	**-0.90**	91%	5%	0%	3%	*1.85**
$9, $1	"Give $1"	**-0.64**	61%	31%	3%	6%	"Take $4"	**-0.83**	82%	14%	1%	3%	*3.13****
$8, $2	"Give $2"	**-0.44**	35%	51%	10%	4%	"Take $3"	**-0.67**	55%	40%	3%	1%	*3.27****
$7, $3	"Give $3"	**-0.16**	8%	62%	26%	4%	"Take $2"	**-0.38**	28%	53%	16%	2%	*3.34****
$6, $4	"Give $4"	**0.14**	3%	30%	61%	7%	"Take $1"	**-0.09**	12%	46%	36%	7%	*3.42****
$5, $5	"Give $5"	**0.87**	0%	3%	14%	83%	"Give $0" / "Take $0"	**0.93**	0%	0%	11%	89%	*1.26*
$4, $6	"Give $6"	**0.57**	0%	7%	50%	43%	"Give $1"	**0.48**	4%	12%	40%	43%	*0.72*
$3, $7	"Give $7"	**0.42**	1%	22%	39%	37%	"Give $2"	**0.31**	7%	23%	38%	33%	*1.12*
$2, $8	"Give $8"	**0.32**	6%	31%	23%	40%	"Give $3"	**0.20**	14%	27%	23%	36%	*1.08*
$1, $9	"Give $9"	**0.22**	17%	24%	19%	40%	"Give $4"	**0.10**	27%	16%	21%	31%	*0.99*
$0, $10	"Give $10"	**0.18**	26%	13%	18%	43%	"Give $5"	**0.04**	36%	10%	16%	38%	*1.13*

* - $p < 0.1$, ** - $p < 0.05$, *** - $p < 0.01$; all two-tailed

Responses are: "very socially inappropriate" (- -), "somewhat socially inappropriate" (-), "somewhat socially appropriate" (+), "very socially appropriate" (+ +); modal response are shaded. To construct the mean ratings, we converted responses into numerical scores ("very socially inappropriate" = -1, "somewhat socially inappropriate" = -1/3, "somewhat socially appropriate" = 1/3, "very socially appropriate" = 1).

Table 2. Statistical tests of behavior across bully vs. standard treatments (data from Experiment 2)

	(1)	(2)	(3)
Dependent variable:	Amount allocated to recipient	Binary (Share = $5)	Binary (Share = $0)
Bully	0.678** (0.210)	1.570*** (0.390)	0.532** (0.248)
Class size	-0.011* (0.006)	-0.018*** (0.004)	-0.002 (0.004)
Constant		0.585 (0.654)	0.091 (0.536)
N	106	97	71
Model:	*Ordered logistic regression*	*Logistic regression*	*Logistic regression*
Sample:	*All data*	*Subjects who allocated less than $6 to recipient*	*Subjects who allocated less than $5 to recipient*

* - $p < 0.1$, ** - $p < 0.05$, *** - $p < 0.01$; all two-tailed
Standard errors (clustered by session) are in parentheses.

Table 3. Conditional (fixed-effects) logit estimation of choice determinants across experiments (includes mean appropriateness ratings from Experiment 1 as an explanatory variable)

Behavioral data (experimental treatment)	Experiment 2 (Standard vs. Bully)		Lazear, et al. (2012) (Standard vs. Sorting)		List (2007) (Standard vs. Take $1)		Data from all three experiments	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Monetary Payoff (β)	0.656*** (0.132)	0.630*** (0.138)	0.811*** (0.075)	0.810*** (0.075)	1.456*** (0.408)	1.312*** (0.401)	0.750*** (0.060)	0.808*** (0.105)
Appropriateness rating (γ)	1.858*** (0.410)	1.556*** (0.521)	2.304*** (0.287)	2.283*** (0.312)	1.941** (0.921)	1.982** (0.843)	1.856*** (0.204)	2.192*** (0.326)
Appropriateness rating X non-standard treatment		0.374 (0.326)		0.062 (0.331)		-0.629 (0.593)		
Monetary payoff X Lazear, et al., experiment								-0.094 (0.127)
Appropriateness rating X Lazear, et al., experiment								-0.125 (0.470)
Monetary payoff X List experiment								0.426 (0.391)
Appropriateness rating X List experiment								-1.029 (1.038)
2γ/β	*5.66*** (0.49)*	*4.94*** (0.98)*	*5.68*** (0.39)*	*5.64*** (0.48)*	*2.67*** (0.98)*	*3.02*** (0.90)*	*4.95*** (0.29)*	*5.43*** (0.30)*
Log-likelihood	-208.5	-207.7	-308.8	-308.7	-126.8	-126.1	-672.3	-649.8
Obs. (subjects)	1166 (106)	1166 (106)	2105 (183)	2015 (183)	816 (70)	816 (70)	4087 (359)	4087 (359)

* - p < 0.1, ** - p < 0.05, *** - p < 0. 01; all two-tailed

Bootstrapped standard errors are in parentheses. The variable "appropriateness rating" converts subject responses in Experiment 1 to numerical scores ("very socially inappropriate" = -1, "somewhat socially inappropriate" = -1/3, "somewhat socially appropriate" = 1/3, "very socially appropriate" = 1).

Figure 1A. Distributions of amounts shared in standard vs. bully treatments (data from Experiment 2)

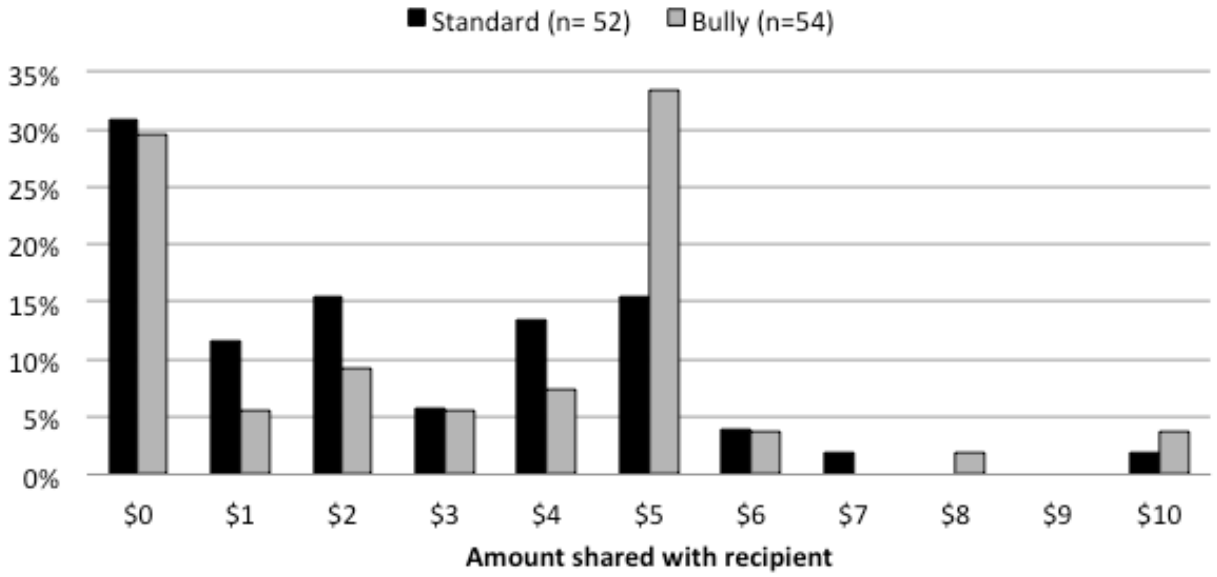

Figure 1B. Predicted distributions of amounts shared in standard vs. bully treatments (based on coefficients in Table 3, Model 1)

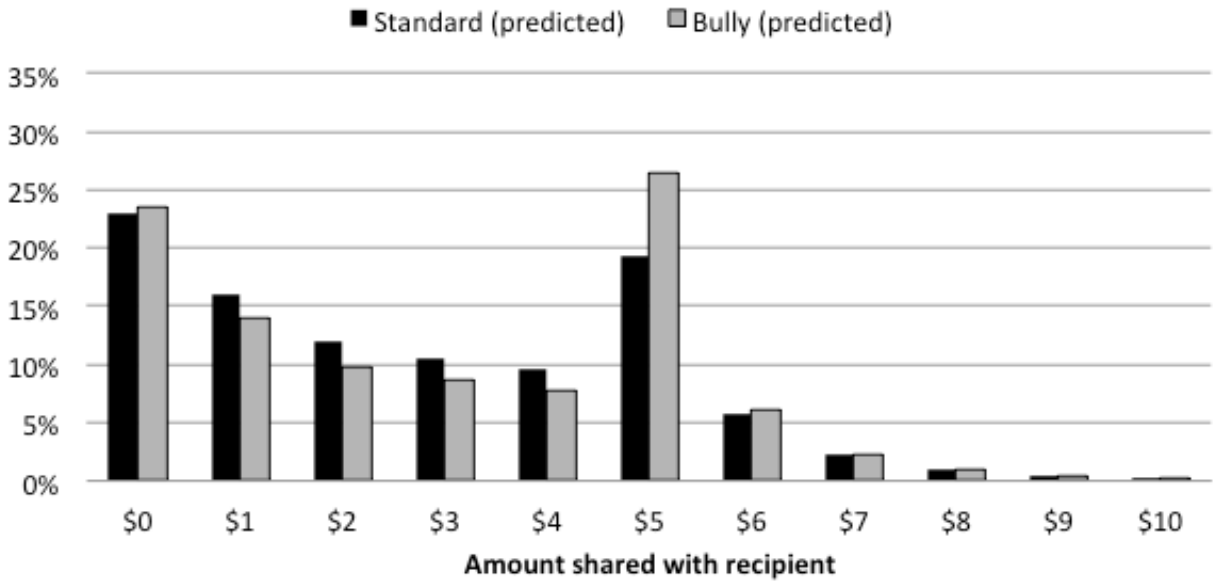

Figure 2A. Distributions of amounts shared in standard vs. sorting treatments (data from Experiments 1 and 2 of Lazear et al. 2012)

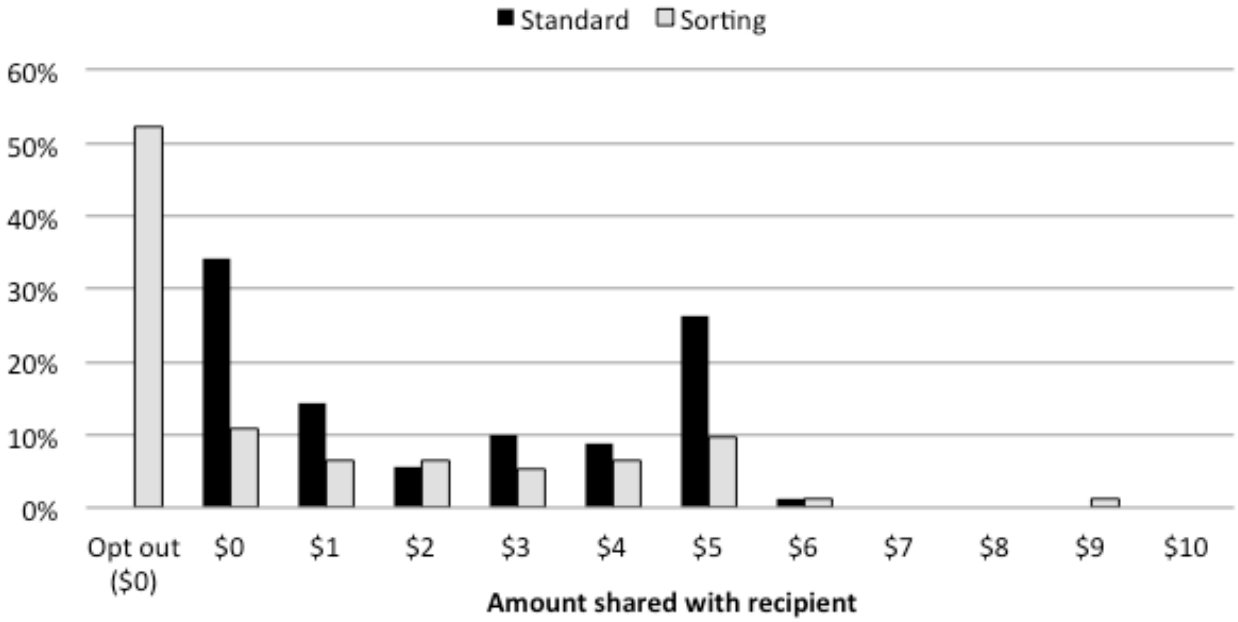

Figure 2B. Predicted distributions of amounts shared in standard vs. sorting treatments (based on coefficients in Table 3, Model 3)

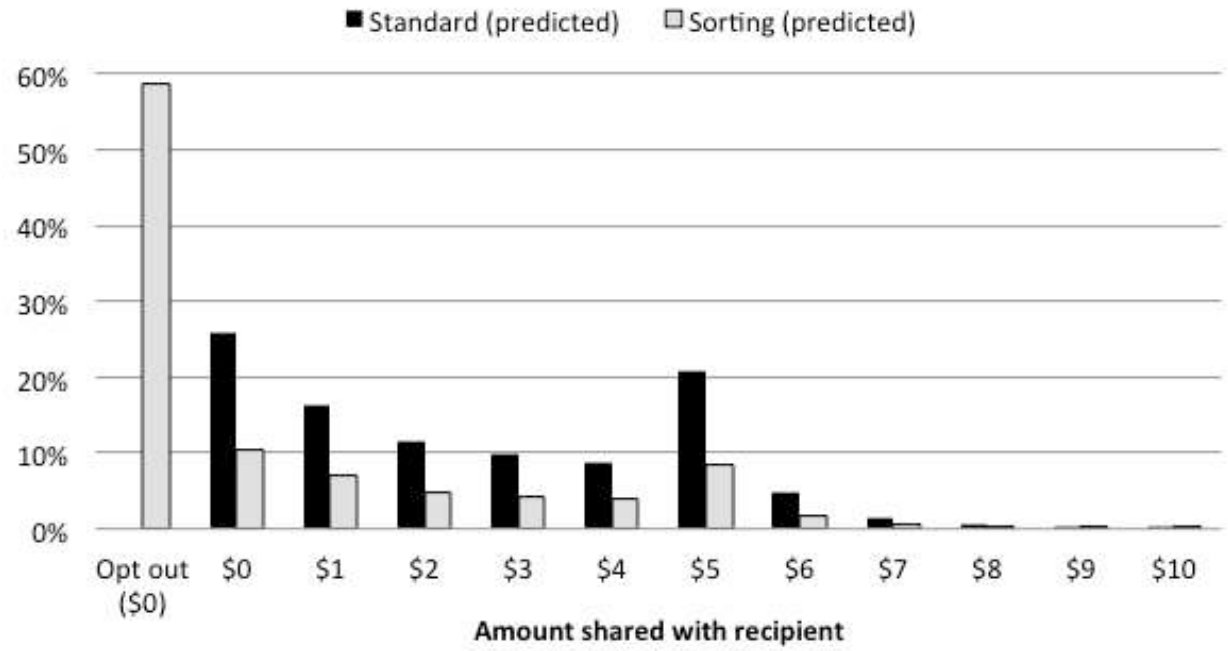

Figure 3. Mean ratings of social appropriateness from standard vs. sorting treatments (data from Experiment 1)

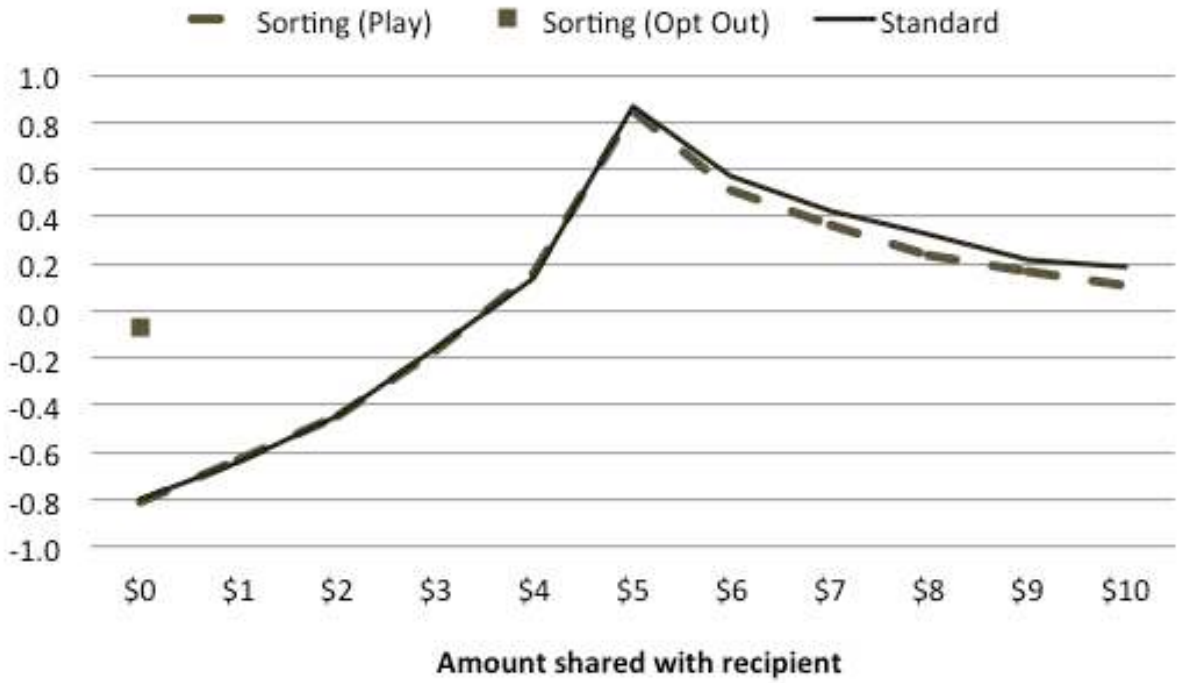

Figure 4A. Distributions of amounts shared in standard vs. take $1 treatments (data from List 2007)

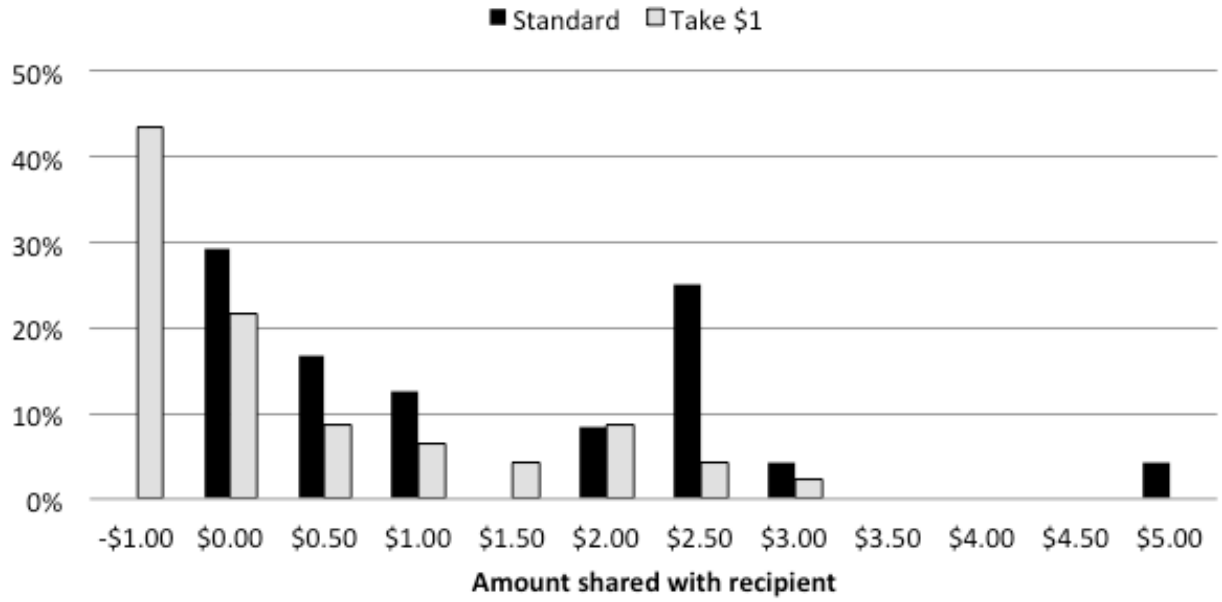

Figure 4B. Predicted distributions of amounts shared in standard vs. take $1 treatments (based on coefficients in Table 3, Model 5)

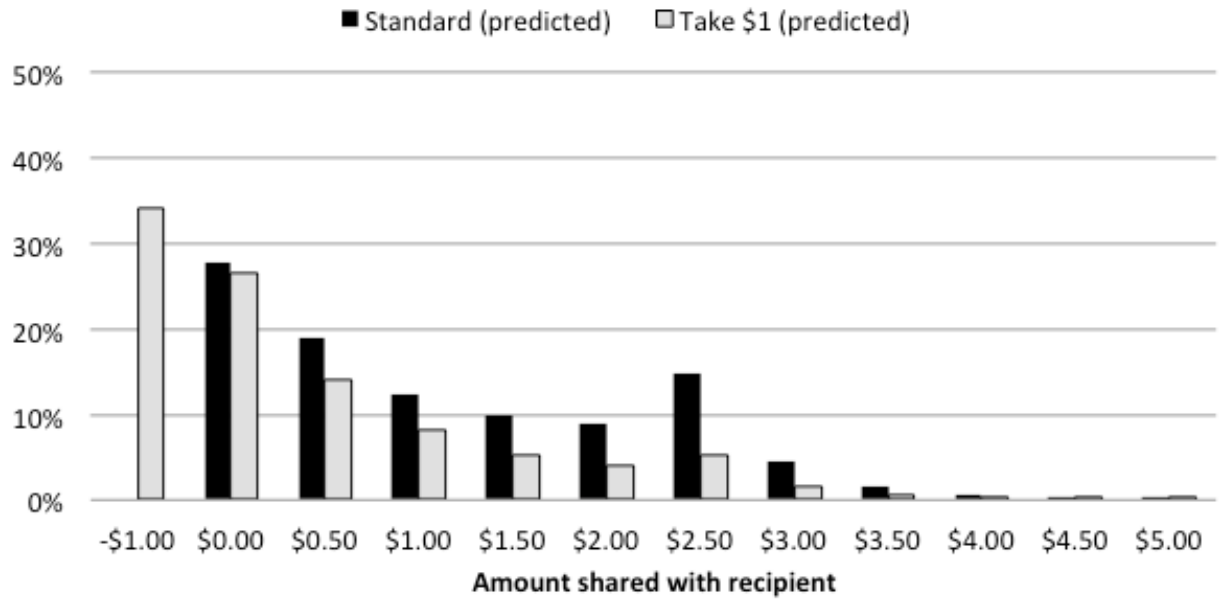

Figure 5. Mean ratings of social appropriateness from standard vs. take $1 treatments (data from Experiment 1)